REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* OCTOBER 2013	**2. REPORT TYPE** Conference Paper	**3. DATES COVERED** *(From - To)* APR 2011 – JUN 2013

4. TITLE AND SUBTITLE

Enhancing the Classification Accuracy of IP Geolocation
(POST PRINT)

5a. CONTRACT NUMBER
IN-HOUSE GGIHZORR

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER
GGIH

5e. TASK NUMBER
ZO

5f. WORK UNIT NUMBER
RR

6. AUTHOR(S)

Tennessee State University: Hellen Mazikuk, Sachin Shetty, Tamara Rogers

Air Force Research Laboratory: Keesook J. Han

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Tennessee State University, 3500 John Merritt Blvd, Nashville, TN 37209

Air Force Research Laboratory/RIGA 525 Brooks Rd, Rome NY 13440

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory/Information Directorate
Rome Research Site/RIGA
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TP-2013-059

12. DISTRIBUTION AVAILABILITY STATEMENT

Distribution Approved For Public Release; Distribution Unlimited.
PA Case number: 88ABW-2012-4255, dated 3 Aug 2012

13. SUPPLEMENTARY NOTES

© IEEE 2013. This paper was published in the Proceedings of the IEEE MILCOM 2012, 29 Oct – 1 Nov 2012, Orlando, FL. This work is copyrighted. One or more of the authors is a U.S. Government employee working within the scope of their Government job; therefore, the U.S. Government is joint owner of the work and has the right to copy, distribute, and use the work. All other rights are reserved by the copyright owner.

14. ABSTRACT

The ability to localize Internet hosts is appealing for a range of applications from online advertising to localizing cyber attacks. Recently, measurement-based approaches have been proposed to accurately identify the location of Internet hosts. These approaches typically produce erroneous results due to measurement errors. In this paper, we propose an Enhanced Learning Classifier approach for estimating the geolocation of Internet hosts with increased accuracy. Our approach extends an existing machine learning based approach by extracting six features from network measurements and implementing a new landmark selection policy. These enhancements allow us to mitigate problems with measurement errors and reduces average error distance in estimating location of Internet hosts. To demonstrate the accuracy of our approach, we evaluate the performance on network routers using ping measurements from PlanetLab nodes with known geographic placement. Our results demonstrate that our approach improves average accuracy by geolocating internet hosts 100 miles closer to the true geographic location versus prior measurement-based approaches.

15. SUBJECT TERMS

IP Geolocation, Machine Learning-based Classification, Network Measurement, PlanetLab, Cloud Computing, Security, Privacy

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** KEESOOK J. HAN
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	7	**19b. TELEPHONE NUMBER (*Include area code*)** N/A

Enhancing the Classification Accuracy of IP Geolocation

Hellen Maziku*, Sachin Shetty*, Keesook Han† and Tamara Rogers*
* College of Engineering
Tennessee State University, Nashville, TN, USA
Email: hmaziku@my.tnstate.edu, sshetty@tnstate.edu, trogers3@tnstate.edu
† Air Force Research Laboratory
Rome, NY, USA,
Email: keesook.han@rl.af.mil

***Abstract*—The ability to localize Internet hosts is appealing for a range of applications from online advertising to localizing cyber attacks. Recently, measurement-based approaches have been proposed to accurately identify the location of Internet hosts. These approaches typically produce erroneous results due to measurement errors. In this paper, we propose an Enhanced Learning Classifier approach for estimating the geolocation of Internet hosts with increased accuracy. Our approach extends an exisiting machine learning based approach by extracting six features from network measurements and implementing a new landmark selection policy. These enhancements allow us to mitigate problems with measurement errors and reduces average error distance in estimating location of Internet hosts. To demonstrate the accuracy of our approach, we evaluate the performance on network routers using ping measurements from PlanetLab nodes with known geographic placement. Our results demonstrate that our approach improves average accuracy by geolocating internet hosts 100 miles closer to the true geographic location versus prior measurement-based approaches.**

I. Introduction

The ability to accurately identify the geographic location of Internet devices has signficant implications for online-advertisers, application developers, network operators and network security analysts. For instance, IP geolocation is used for enforcing digital content and territory rights (e.g., Pandora, BBC Iplayer, Real Media, Comedy Central, Netflix and Spotify) and target advertising (e.g., Google). More recently, IP geolocation techniques have been deployed in cloud infrastructure services. For example, a Dropbox user may require their data to be hosted on servers in San Francisco, but the data's true location may actually be in Tennessee (see Fig. 1). Users of cloud computing deploy Virtual Machines (VM) on a cloud providers infrastructure without having to maintain the hardware their VM is running on. However, cloud auditing policy requires that cloud providers restrict VM locations to certain datacenters, as specified by a Service Level Agreement (SLA). Cloud users can use IP geolocation to independently verify data confidentiality by ensuring location restrictions in their cloud SLAs are met.

The problem of estimating geographical location of hosts on the Internet by a single IP address presents several challenges due to lack of a relationship between the IP address and its geographic location. Fig. 2 provides an overview of current IP geolocation approaches. Over the last decade, several IP geolocation approaches aimed at accurate approximation of the location of network hosts have emerged. These approaches can be broadly classified into two groups depending on their technique to collect location information. One set of techniques leverages information from commercial databases to procure information on the geographic location of IP addresses. These databases store organizational information assigned to IP domains and DNS names. Databases such as ARIN [13] and QUOVA [12] utilize previously registered data to geographically locate an IP address. These databases tend to be coarse grained, usually returning the headquarters location of the organization that registered the IP address. This becomes a problem when organizations distribute their IP addresses over a wide geographic region, such as large ISPs or content providers. The databases can also be easily fooled by proxies.

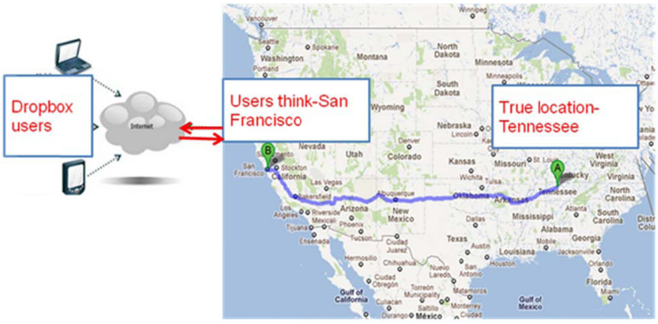

Fig. 1. IP geolocation may be used by users to verify the true location of their data

The other technique utilizes active delay and topology measurements to estimate the geographic locations of IP addresses. Measurement-based geolocation approaches which use end-to-end RTTs are classified as delay-based algorithms and those that use both RTT and topology information as topology-aware algorithms. But topology based approaches

Approved for Public Release; Distribution Unlimited: 88ABW-2012-4255, 3 AUG 12

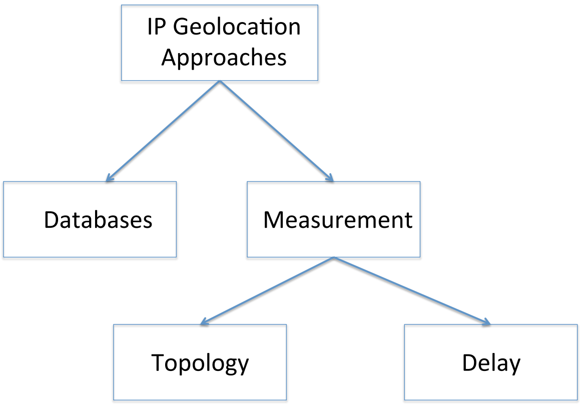

Fig. 2. IP geolocation approaches

are typically plagued by inaccuracies in geolocating nodes. A recent study [18] reveals that topology-aware approaches actually fare worse against an adversary who tries to subvert the approaches into returning forged results. Topology-aware approaches are therefore less suitable for security-sensitive applications.

This paper presents a Enhanced Learning Classifier IP geolocation approach, which extends an existing machine learning based IP geolocation aprroach with additional features and proper landmark selection. Recently, it has been shown that accuracy of IP geolocation can be improved by casting IP geolocation as a machine learning classification problem. This approach makes it possible to incorporate both network measurements (latency and hop count) and societal characteristics (city population density) to locate an IP address. We improve the accuracy of the existing machine learning IP geolocation classifier by expanding the list of features to include average delay, standard deviation of delay, mode and median of delay and careful selection of landmarks. The addition of these features ensure that the our approach is less prone to measurement errors and other network anomalies affecting the distance approximation. To demonstrate the robustness and the accuracy of our approach, we evaluate the performance on PlanetLab nodes. Our performance analysis shows that on ground truth data sets, our approach provides location estimations with outstanding accuracy compared to state-of-the-art techniques, Constraint Based Geolocation (CBG) [15] and Machine Learning Based IP Geolocation [3].

The rest of the paper is organized as follows. In Section II we provide an overview of state-of-the art active geolocation approaches. In Section III we present our Enhanced Learning Classifier to geolocate Internet hosts. In Section IV we provide the performance evaluation of our approach by examining the accuracy of location estimations. We finally conclude in Section V.

II. Related work

Our proposed IP geolocation approach can be categorized under the measurement-based approach. We compared the performance of our approach with CBG [15] and Machine Learning based IP Geolocation [3]. In most measurement-based approaches, we have landmark nodes with known geographic location and a target node without known position. To estimate the location of the target we measure network delays from the landmarks to the target, and then convert the delays into geographic distances based on a delay-distance function. These delay based approaches only differ in how they express the distance to delay function and how they triangulate the position of the target.

Statistical Geolocation [14] develops a joint probability density function of distance and delay that is input into a force-directed algorithm used to geolocate the target. CBG [15], on the other hand, establishes the distance-delay function, by having the landmarks ping each other to derive a set of points mapping geographic distance to network delay. Each landmark then computes a linear (best fit) function that is closest to, but below, the set of points. The distance between each landmark and the target IP is inferred using the best fit function, creating an implied circle around each landmark where the target IP may be located. The target IP is then predicted to be in the region of intersection of the circles of all the landmarks. Spotter [17] derives a common delay-distance model using a probabilistic approach based on a detailed statistical analysis of the relationship between network delay and geographic distance. The delay-based approach assumes that network delay is well correlated with geographic distance. However, network delay is composed of queuing, processing, transmission and propagation delay, where propagation delay is related to distance traveled, and the other delay components vary depending on network load, thus adding noise to the measured delay. This assumption is also violated when network traffic does not take a direct path between hosts. Therefore, delay-based approaches that depend on only delay measurements as an input to their geolocation frameworks continue to produce low accuracy results. One of the the most recent delay-based IP geolocation approach [3] models IP geolocation as a machine learning classification problem. This classification-based approach makes it possible to incorporate other types of geolocation information into the framework other tha relying on network delay alone.

III. Enhancing Learning Classifier

The goal of our approach is to improve the accuracy of geolocating internet hosts by selecting appropriate network measurement features and choosing landmarks based on maximizing coverage and responsiveness. To that end, we identified 6 features that our classifier uses in estimating location of internet hosts. Of these features, 3 have not been proposed before in previous research. We performed a passive analysis of network traffic generated by probes sent from landmarks to targets. We identified features that were able to reduce measurement errors in estimating network delay between landmarks and targets.

A. Overview of Machine Learning approach

Erikkson et.al., proposed the machine learning IP geolocation approach that uses delay and hop measurements to geolocate network routers [3]. This approach employs a Naive Bayes framework to convert network measurements betweeen landmarks and network routers to distances.

Let the network measurements from j landmarks to single target be recorded as $M = \{m_1, m_2, ..., m_j\}$. Using Bayes theorem, Eriksson et al. [3] estimates the county ($\hat{c}$) of the target IP address as;

$$\begin{aligned}\hat{c} &= \arg\max_{c\in C} p(c|M) \\ &= \arg\max_{c\in C} \tfrac{p(M|c)p(c)}{p(M)} = \arg\max_{c\in C} p(M|c)p(c)\end{aligned} \tag{1}$$

where $p(c)$ is the probability of classifiying target in county c and $p(M|c)$ is the conditional probability of M being observed given target county c. The value $p(M)$ is the probability of observing the measurement set can be ignored due to this value being constant for any chosen county c.

Other than measurements from landmarks to IP targets, Eriksson et al. also incorporate another feature, namely county population density in their classifier. The county population density for United States is publicly available on the U.S. Census Bureau website [19]. The $p(c)$ value is chosen based on the fact that the number of routers in a specific geographic location is strongly correlated with the population of that geographic location [1].

The $p(M|c)$value is estimated using kernel density estimators [6]. Assuming the entries of M are statistically independent [11], we have;

$$\begin{aligned}p(M|c) &= p(\{m_1, m_2, ..., m_j\}|c) \\ &= p(m_1|c)p(m_2|c)...p(m_j|c)\end{aligned} \tag{2}$$

The aim is now to determine $p(m_i|c)$ for each of j landmark locations and later substitute these values in equation (2). With known j landmark locations, and known target location, we can get the true distances from all these landmarks to the target as $d = \{d_1, d_2, ..., d_j\}$. These true distances can be used to learn the density (the probability of observing measurements, m_i given that the target is located in d_i distance away from the i-th landmark. Casting IP geolocation as a machine learning-based classification problem enables information from multiple datasets to be fused such that areas that have low information content from one measurement can be compensated with better information content from other measurements.

B. Feature Selection

To determine the measurement features that model network delay accurately, we analyzed the network delay collected from probes generated between various landmark and targets for one month. Table1 gives an overview of the measurement features that we identified. In this section, we describe these features and explain why we believe that they can contribute towards increased geolocation accuracy.

TABLE I
FEATURE LIST

k	1	2	3	4	5	6
Features	Avg	Hop	Mode	Median	Std. Dev.	Pop Density

Avg (Average Delay), Hop (Hop Count), Mode (Mode of Delay), Median (Median of Delay), Std. Dev. (Standard Deviation of Delay) and Pop. Density (Population Density)

The existing learning based IP geolocation approach uses two measurement features (delay and hop count) and one societal feature (population density). But the approach suffers from large estimation errors that are caused by imperfect measurements, sparse measurement availability, and irregular Internet paths. In our approach, we mitigate these issues by expanding the feature set. Our approach introduces three new network measurement features (standard deviation of delay, mode of delay and median of delay). Every target exhibits different statistical behavior due to variations in the underlying network between target and landmark. Instantaneous delay measurements alone cannot capture the statistical variations of the targets. For instance, probes between landmarks and targets usually do not traverse the same network path. Each network path introduces new routers which result in variation of network delay. By incorporating the variance in the network delay as features, we can improve the accuracy of the classifier. The addition of the three new measurement features comprehensively address all statistical variations of delay from landmarks to targets.

Assume targets *x* and *y* each receive *n* ICMP echo (ping) requests sent from a single landmark. Average delay is used as a good estimate in predicting subsequent data points. If deviations in each of the *n* ICMP echo requests from the landmark to the *x* target are very minimal, then mean delay will be a good representative to capture the information encoded by *n* ICMP echo requests. But due to network dynamics, it is quite possible that there exist sufficient deviation in the *n* values from the landmark to the target *y*. The deviations may not be uniform and may change with time. We need additional features to capture this impact of network dynamics on the delay measurements. We add median and mode features to gain an accurate estimate of the “middle” of a set of latency data. Finally, we add standard deviation to the list of features to estimate the amount of deviation from the average delay. Table I provides an overview of the features used in the Enhanced Learning Classifier approach.

Our measurement set $M = \{m_1, m_2, ..., m_j\}$, where $m_j = \{m_{jk}\}$ and $k = 1, 2, 3, 4, 5, 6$ (where the total number of measurements to the target IP address is given by $M = 6j$)

A set of nonnegative estimation weights $\{\lambda_k\}$ is introduced for each feature to reflect the importance and contribution of that feature in the overall classification process [3]. Another set of estimation weights $\{\gamma_k\}$ is introduced to order the landmarks. A landmark with the smallest feature measurement values weighs most and informs the classifier more accurately than a landmark with the largest values. The weight parameter values will be chosen by the least squares parameter estimation method. The method minimized the sum of squared distance errors between the training set of IPs known locations and the estimated locations.

With the features and the estimation weights in place, the classifier in equation (1) may be restated in logarithm as;

$$\hat{c} = \arg\max_{c\in C}(\textstyle\sum_{k=1}^{5} \lambda_k \sum_{j=1}^{J} exp(-j.\gamma_k) log\ p(m_{jk}) + \lambda_6 log\ p(c)) \quad (3)$$

where J is the total number of measurements from landmarks to target IP.

Procedure 1 Enhanced Learning Classifier

```
1: Identify set of landmarks based on satifying coverage and responsiveness criteria.
2: Collect network measurements (hop count and latency) from each landmark to a set of nodes with known geographic locations.
3: Use census database to estimate prior probability.
4: Expand feature selection to include average, mode, median, and standard deviation of latency measurements.
5: Perform kernel density estimation to estimate one-dimensional distribution for each of the six features.
6: Find optimal values for each of the six features to minimize sum of squared distance errors over the training set.
7: For each IP address with unknown geography, estimate location using equation (3).
8: Perform 5-fold Cross Validation
9: Compute error distance by comparing with ground truth database provided by Maxmind
```

IV. Performance Evaluation

A. *Identifying target IP addresses*

To evaluate our approach, we present the router localization scenario, which is critical for examining the geographical properties of the Internet. In this scenario, we collected as many spatially diverse router IP addresses as possible within continental United States. We identified routers along multiple network paths between all PlanetLab node pairs. To collect the router IP addresses, we performed a full mesh traceroute between responsive PlanetLab nodes multiple times between June 1, 2011 and October 31, 2011. The traceroute probes resulted in 142,937 router IP addresses. We used Maximind [7] database to filter out the IP addresses that fall within the United States and have known city and county locations. As different IP addresses could potentially belong to unique interfaces of the same router, we performed IP dealiasing to reduce the router IP addresses to 23,843.

B. *Identifying landmarks*

The selection of landmarks plays a critical role in improving the accuracy of measurement based IP geolocation. Therefore, identifying suitable landmarks to reduce error distances is a crucial process. But a standard protocol or standard procedure that outlines a clear approach to landmark selection is lacking. Prior methods have not presented techniques and motivations for selecting landmarks. Landmark selection remains an ambiguous task.

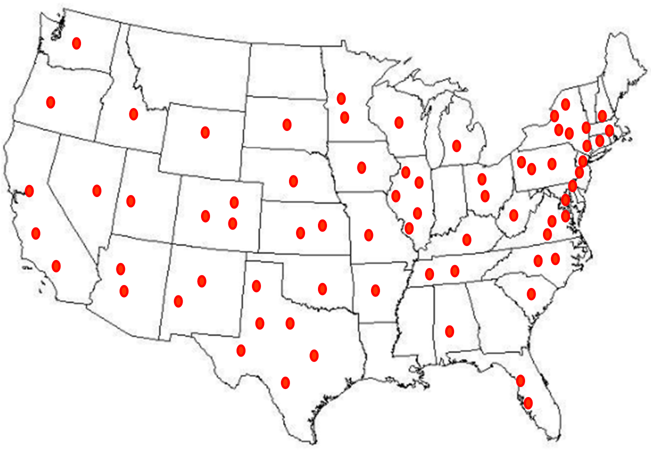

Fig. 3. The distribution of our landmarks

To identify our landmarks, we use the CoMon [2] project of PlanetLab to retrieve a list of the active PlanetLab nodes. Out of 1090 available PlanetLab nodes, 860 nodes responded to our measurement probes. We used Maximind to filter the alive PlanetLab nodes in the United States. As a result of the filtering process, we found 308 alive PlanetLab nodes in the United States. Ethan KatzBassett et al. [4] discovered that distance from a target to the nearest landmark strongly predicts the estimation error. Therefore, delay-based techniques only provide consistent quality if landmarks are ubiquitous. To complement this discovery, we select a few targets in New York state, as well as all the landmarks in New York state and perform IP geolocation. We then use latitudes and longitudes of the landmarks to eliminate those that are within the same city, leaving out only one landmark per city. We perform geolocation using the second set of filtered landmarks with identical results. This process demonstrates that even though landmarks need to be ubiquitous for consistent quality, concentrating many of the landmarks in the location relative to a target does not improve the IP geolocation reults of that target. With these findings, we reduce the set of 308 landmarks to 108. Out of 108 landmarks, 67 landmarks were able to send an ICMP echo request to the 23,843 targets.

The distribution of our landmarks is limited by the existence or non-existence of PlanetLab nodes in certain areas. Fig. 3 shows our landmark distribution.

To show the impact of landmark distribution on the accuracy of IP geolocation, we divide our testing set into four regions in the United states; North East region, West region, North Central region and South region. Targets in the North East region return an average error distance of 23.88 miles with a median of 0 miles and a maximum error distance of 127 miles. Only 8 out of 151 targets return error distances of

TABLE II
COMPARISON OF THE MEAN ERROR DISTANCES (IN MILES) WITH PRIOR METHODS

	avg	(avg, hop)	(avg, hop, pop density)	(avg, hop, pop density, std)	(avg, hop, pop density, std, mod)
CBG	322.49				
Learning-Based	278.96	261.89	253.34	-	-
Enhanced Learning Classifier	270.35	216.80	206.55	176.33	155.74

avg (average delay), hop (hop count), pop density (population density), std (standard deviation of delay) and mod (mode of delay)

more than 100 miles. The rich concentration of landmarks in the North East region account for the excellent results, which is not the same case in the other regions. Therefore, delay-based approaches only provide consistent quality if landmarks are ubiquitous.

C. Collect Measurements

To generate the measurement set, we collect instantaneous delay and hop count measurement from each of the 23,843 targets to the 67 landmarks. For the instantaneous delay data, we send 40 ICMP echo requets from each landmark to all the targets. Based on the instantaneous delay measurements, we calculate the average, standard deviation, mode and median of delay for each target from each landmark, which results in $67 \times 23,843 \times 4 = 6,389,924$ measurements.

Using traceroute to collect hop count measurements causes excessive overhead on the core routers. To avoid this overhead we send a single ICMP echo request from each landmark to all targets. We then use this request to calculate the hop count of the reverse path [3,5]. We use 5-Fold Cross Validation to test the performance of the methodology five times using 80% of the routers as our training set, leaving the remaining 20% of the routers for testing our classifier.

D. Preliminary results

Fig. 4 plots the average error distances in miles with respect to the empirical cumulative probability for the five features. This plot signifies the hypothesis that the accuracy of IP geolocation can be improved by feature expansion, both network and environmental features. Table II compares our preliminary results with learning based IP geolocation and CBG. As see in the table, the average error distance estimates produced by our technique is lesser than learning-based IP geolocation and CBG. Even with same set of features as learning-based IP geolocation, the geolocation estimates produced by our technique is better due to our landmarks selection policy.

E. PlanetLab Challenges

We faced two key challenges while using PlanetLab as a network measurement platform: unresponsive PlanetLab nodes and coverage of PlanetLab nodes. As described in the prior sections, 38% of our total landmarks were unresponsive to ICMP probes. Unresponsive landmarks translate to inability to geolocate targets in those respective areas.

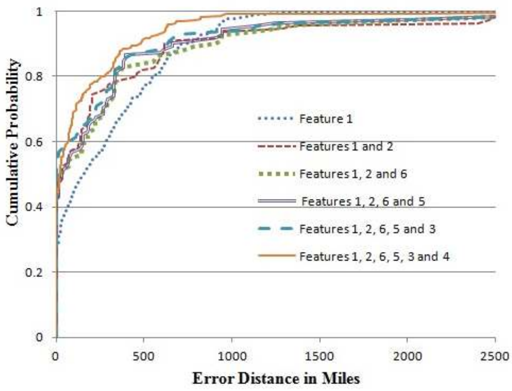

Fig. 4. Empirical cumulative probability of error distance. Features used in each curve are labeled according to Table I

TABLE III
ERROR DISTANCES (MILES) ACROSS FOUR REGIONS IN THE UNITED STATES

	North East	North Central	West	South
Mean error distance	23.88	347.20	231.18	187.84
Median error distance	0	290.72	53.65	95.08
Maximum error distance	127.36	1270.74	882.06	706.01

We do not have landmarks in some of the states, e.g Alaska and Hawaii due to absence of PlanetLab nodes. Targets within these states return error distances of more than 1000 miles. Distance from a target to the nearest landmark strongly predicts the estimation error. Therefore, delay-based techniques only provide consistent quality if landmarks are ubiquitous. The absence of PlanetLab landmarks in some areas remains to be another challenge to delay based IP geolocation techniques.

To illustrate these two challenges in detail, we divide our test dataset (1173 targets) into four regions in the United States: North East region, West region, North Central region and South region. These regions wre identified according to the concentration of our landmarks in the United States (see Fig. 3). The North East region has 23 of the 67 landmarks (34% of total landmarks). Targets in the North East region return an average error distance of 23.88 miles with a median of 0 miles and a maximum error distance of 127 miles. Only 8 out of 151 targets return error distances of more than 100 miles. While targets in the North East region give us excellent results, it is not the same with the rest of the regions. The poor concentration of landmarks, especially in the North Central and West regions highly affects the accuracy of targets in those regions. Table III gives the mean, median and maximum error distances of targets in all the four regions.

V. Conclusion and Future work

In this paper, we proposed an Enhanced Learning Classifier approach to improve the accuracy of estimates of the geographic location of Internet hosts. Our approach extends an existing machine learning based IP geolocation framework by adding new features and careful selection of landmarks based on responsiveness and coverage. The additional features in our approach model the variance in network delay and provide a better statistical description for the relationship between network delay and geographical distance. To demonstrate the accuracy and robustness of our approach, we evaluate the performance on PlanetLab nodes. The results show that the addition of the new features and our landmark selection policy does improve the overall estimation accuracy. For future work, we plan to investigate additional features, both network based and societal based, that lead to accuracy improvement in IP geolocation. Finally, we will perform IP geolocation of routers on GENI's [9, 10] multiple control frameworks to mitigate the challenge of unresponsive landmark nodes.

Acknowledgment

This work was partially supported by National Science Foundation (NSF) Grant HRD-1137466, NSF HBCU-UP Targeted Infusion Grant HRD-1137544, Department of Homeland Security(DHS) SLA grant 2011-ST-062-0000043, 2011-ST-062-0000045. It is also based upon work supported by the AFRL/RI Information Institute VFRP No. R730719.

References

[1] A. Lakhina, J. Byers, M. Crovella, and I. Matta, "On the Geographic Location of Internet Resources," IEEE Journal on Selected Areas in Communications, 2003.

[2] "CoMon-A Monitoring Infrastructure for PlanetLab", http://comon.cs.princeton.edu/

[3] B. Eriksson, P. Barford, J. Sommers, and R. Nowak, "A learning-based approach for IP geolocation," Passive and Active Measurement Workshop, 2010.

[4] K. Ethan, et. al., "Towards IP geolocation using delay and topology measurements," ACM SIGCOMM, 2006.

[5] H. Wang, C. Jin, and K. Shin,"Defense against spoofed IP traffic using hop-count filtering," IEEE/ACM Transactions on Networking, 2007.

[6] L. Wasserman, "All of Nonparametric Statistics," 2007.

[7] Maxmind, 2011. http://www.maxmind.com.

[8] Planetlab, 2010. http://www.planet-lab.org.

[9] The ProtoGENI Control Framework for GENI Cluster C, http://www.protogeni.net/.

[10] A Prototype of a Million Node GENI, http://groups.geni.net/geni/wiki/MillionNodeGENI.

[11] I. Rish, "An Empirical Study of the Naive Bayes Classifer," Workshop on Empirical Methods in Artificial Intelligence, 2001.

[12] Quova IP geolocation experts, 2010. http://www.quova. com

[13] American Registry for Internet numbers (ARIN), 2010. http://www.arin.net.

[14] I. Young, B. Mark, and D. Richards, "Statistical geolocation of Internet hosts," International Conference on Computer Communications and Networks, 2009.

[15] B. Gueye, A. Zivian, M. Crovella, and S. Fdida, "Constraint-based geolocation of Internet hosts," IEEE/ACM Transactions on Networking, 2006.

[16] J. David, "Handbook of Parametric and Nonparametric Statistical Procedures," fifth Edition. Chapman and Hall/CRC, 2011

[17] S. Laki, P. Matray, P. Haga, T. Sebok, I. Csabai, G. Vattay, "Spotter: a model based active geolocation service," IEEE INFOCOM, 2011.

[18] P. Gill, Y. Ganjali, B. Wong, and D. Lie, "Dude, where's that IP? Circumventing measurement-based IP geolocation," USENIX Security Symposium, 2010.

[19] U.S. Census Bureau http://www.census.gov